# 大数据计算基础大作业报告

1190200811 邵超博

## 第一章 绪论

[研究问题的背景]

随着目前数据的爆炸式增长，对于同一类任务有很多不同的算法被研究出来,（例如：分类、回归）但是对于具有不同特点的数据集而言，并不是所有算法最终的性能都很好，算法之间最后的性能差异有很大的浮动。所以，如何根据数据集和算法的特点，为每个数据集尽可能地选择最适合它的算法成为目前的研究热点。

[研究问题的挑战]

问题最大的挑战是从这些数据集和算法中重新提取出新的数据集。这一提取过程十分缓慢，以至于提取时间十分漫长的同时，提取出的数据量还十分少。

[当前研究工作的不足之处]

第一个不足之处在于提取新数据集和训练的模型十分简单，可能存在准确度不高的问题。

第二个不足之处在于为了使运行时间不至于漫长得让人难以忍受，本文使用的数据集都相对比较小。

[本文的工作要解决的问题以及方法]

本文主要解决两大问题，一个是特征提取，即使用不同算法对数据集进行训练和评估并提取出新的数据集，另一个是用新的数据集进行训练并评估训练结果。

本文还试图使用Hadoop加速特征提取的过程。

[本文的贡献]

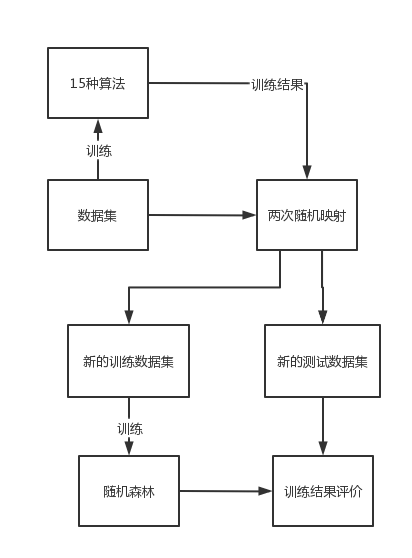
本文初步解决了上面提出的两大问题，使用两次随机投影进行特征提取，使用随机森林进行训练，并使用最优性和准确性两种指标对训练结果进行评价。

本文使用Hadoop加速特征提取的过程，取得了初步的成果。

[章节安排]

本文第二章讲解系统的整体框架。第三章和第四章讲解系统的两个重要部分——特征提取和评价标准。第五章讲解使用Hadoop加速特征提取的方法。第六章进行实验。第七章与第八章介绍相关工作及结论。

## 第二章 系统/方法框架



## 特征提取

将数据集经过两次随机映射得到一个大小为10的double数组，具体过程如下：

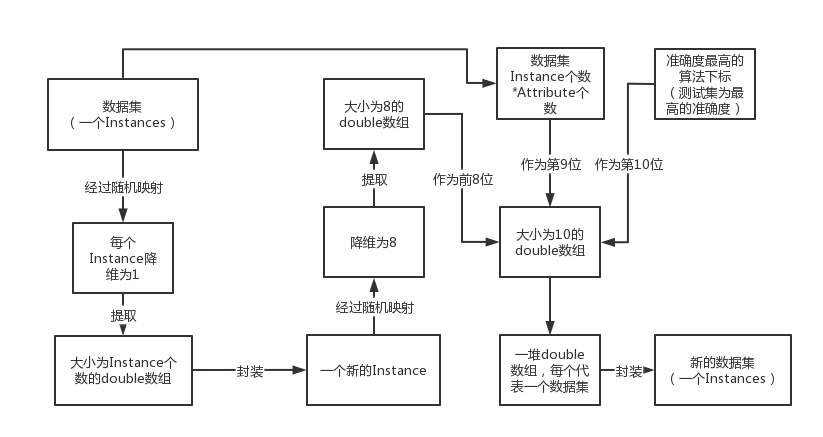
将数据集的每一个Instance经过Random Projection，降维至维数为1。提取每个Instance剩下的一维至一个大小为Instance个数的double数组。

将double数组封装为一个新的Instance，经过一个新的Random Projection，降维至8，提取后得到一个大小为8的double数组。

将每个数据集的Instance个数\*Attribute个数作为double数组的第9位，将准确度最高的算法对应的下标作为double数组的第10位（对于测试数据集，第10位是最高的准确度（accuracy））。第10位也是后续分类算法Random Forest的分类目标。

现在每个数据集都对应一个大小为10的double数组，我们将这些double数组封装为一个Instances，就得到了新的数据集，完成了特征提取的过程。

特征提取的全过程如下：



我们将新的训练数据集扔进Random Forest分类器进行训练，并使用测试数据集评价训练结果。

## 第四章 评价指标

本文使用了两种评价指标，分别是最优性与准确性。

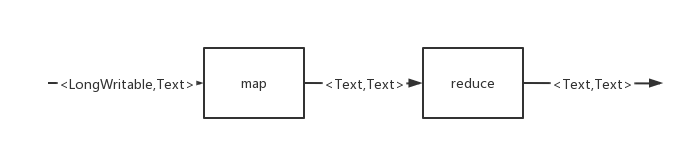
对于单个数据集（原始数据集）来说，最优性是我们给出的算法的准确度与最高的准确度之间的比值；准确性是我们给出的算法的准确度是否等于最高准确度，等于则准确度为1，不等于则准确度为0。即：

对于一堆数据集来说，最优性与准确性都是单个数据集的值的算数平均数。

## 第五章 使用Hadoop对特征提取进行加速

将训练数据集的部分使用Hadoop进行并行化。

Map和Reduce两个函数整体情况如下：



首先将数据集名称和分类器名称以空格分隔按行写入in.txt文件中。Hadoop自动将文件内容读入后，会将整行的内容赋给map函数的value（行号赋给map的key）。

总体来说，map负责进行训练。Map从value获得数据集的名称和分类器的名称后进行训练和评价，获得准确率。将数据集名称作为key，将分类器名称和准确率（以空格分隔）组成的字符串作为value，传递给reduce。

Reduce负责从众多的分类器名称和准确率中找到准确率最高的那一个，并将准确率最高的分类器名称转化为在classifier数组中的下标。将数据集名称作为key，将分类器下标和最高的准确率（以’\t’分隔）组成的字符串作为value，由Hadoop自动输出到文件中。

DataFeature类从文件中读取信息，并进行随机映射（特征提取），生成Instances。

## 第六章 实验

[实验设计]

在使用Hadoop加速的条件下，运行系统的全过程，进行计时。（本地模式的Hadoop）

[对比实验]

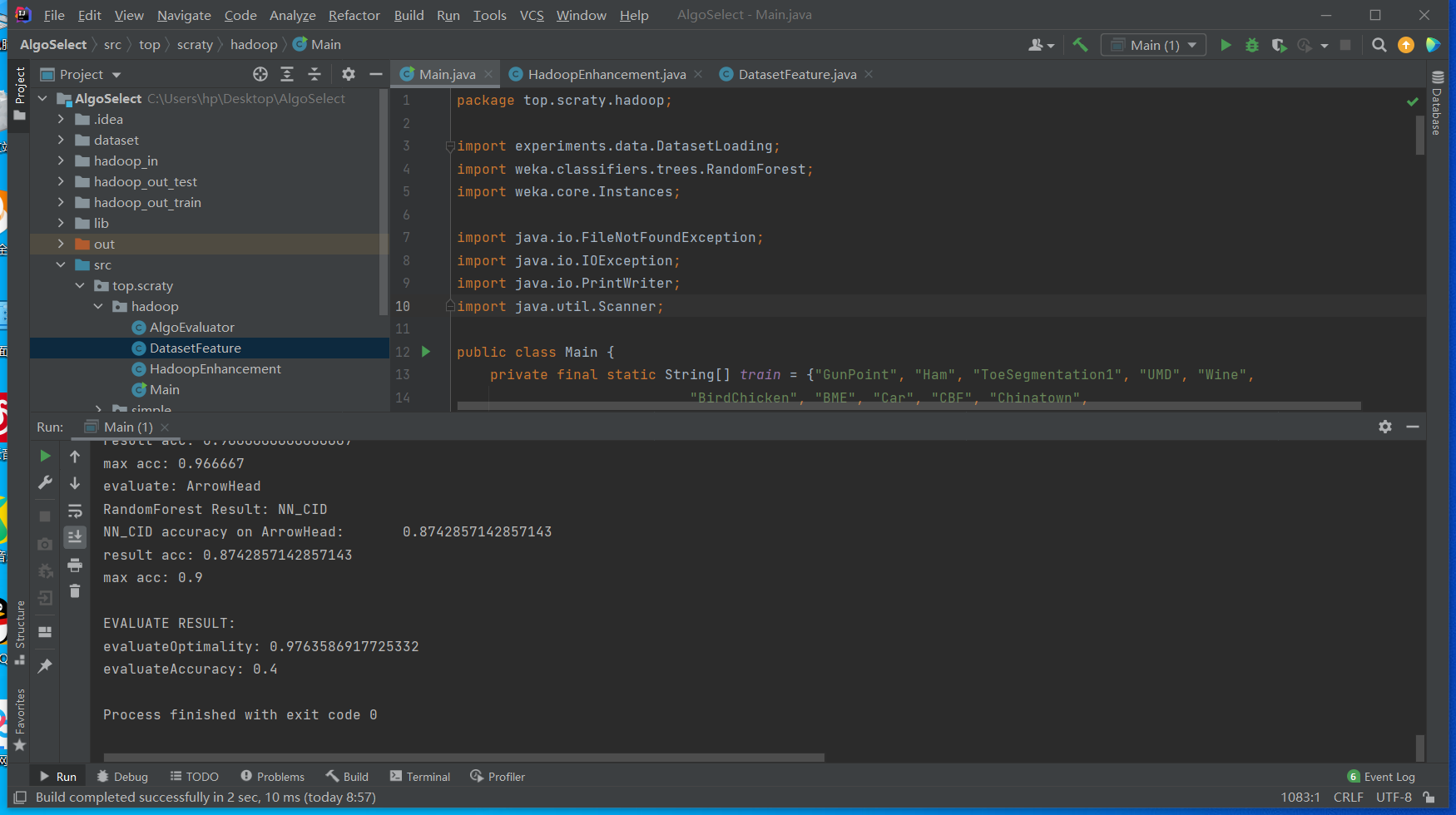
在没有使用Hadoop加速的条件下，运行系统的全过程，进行计时。

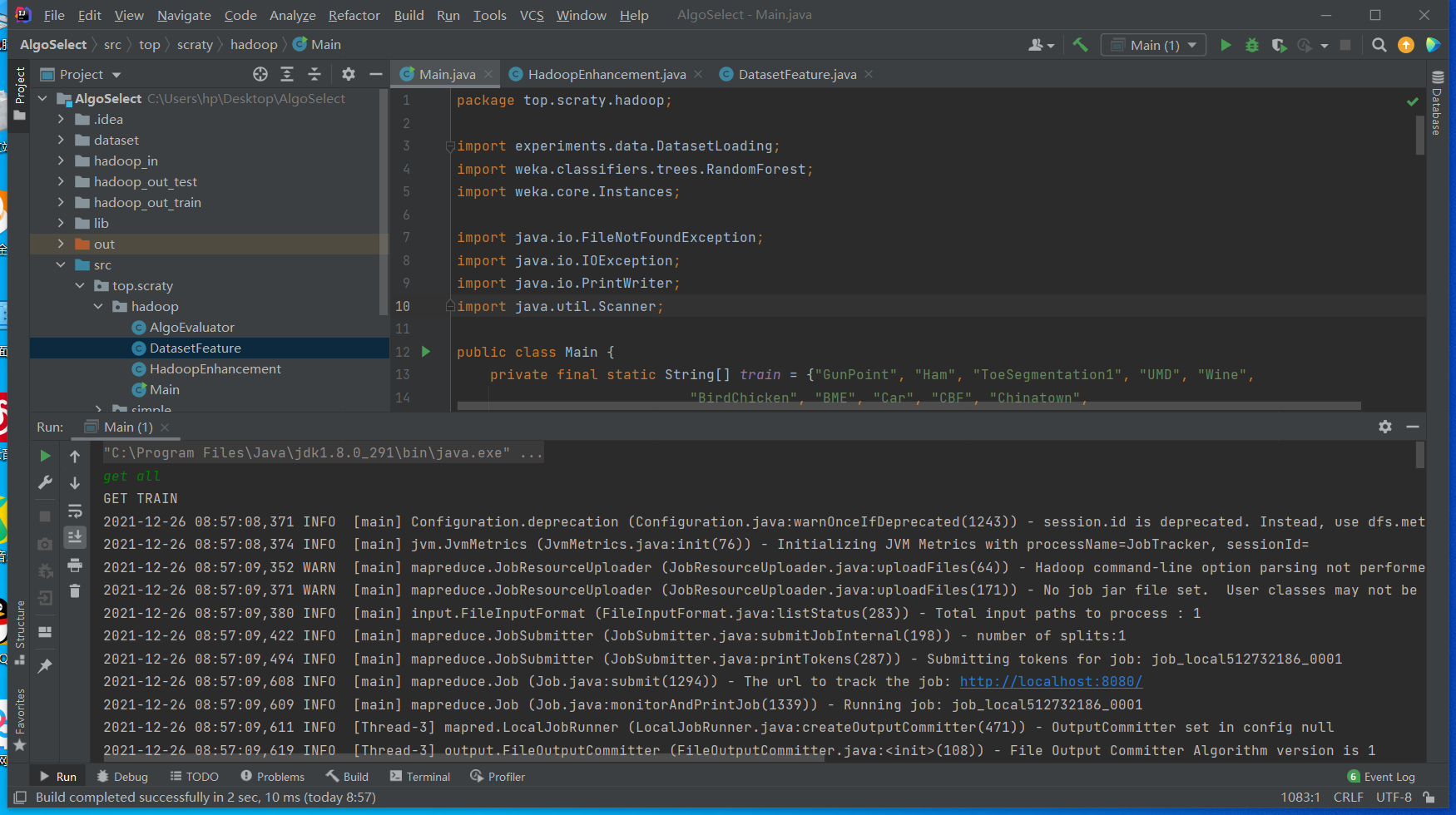
[实验受参数的影响]

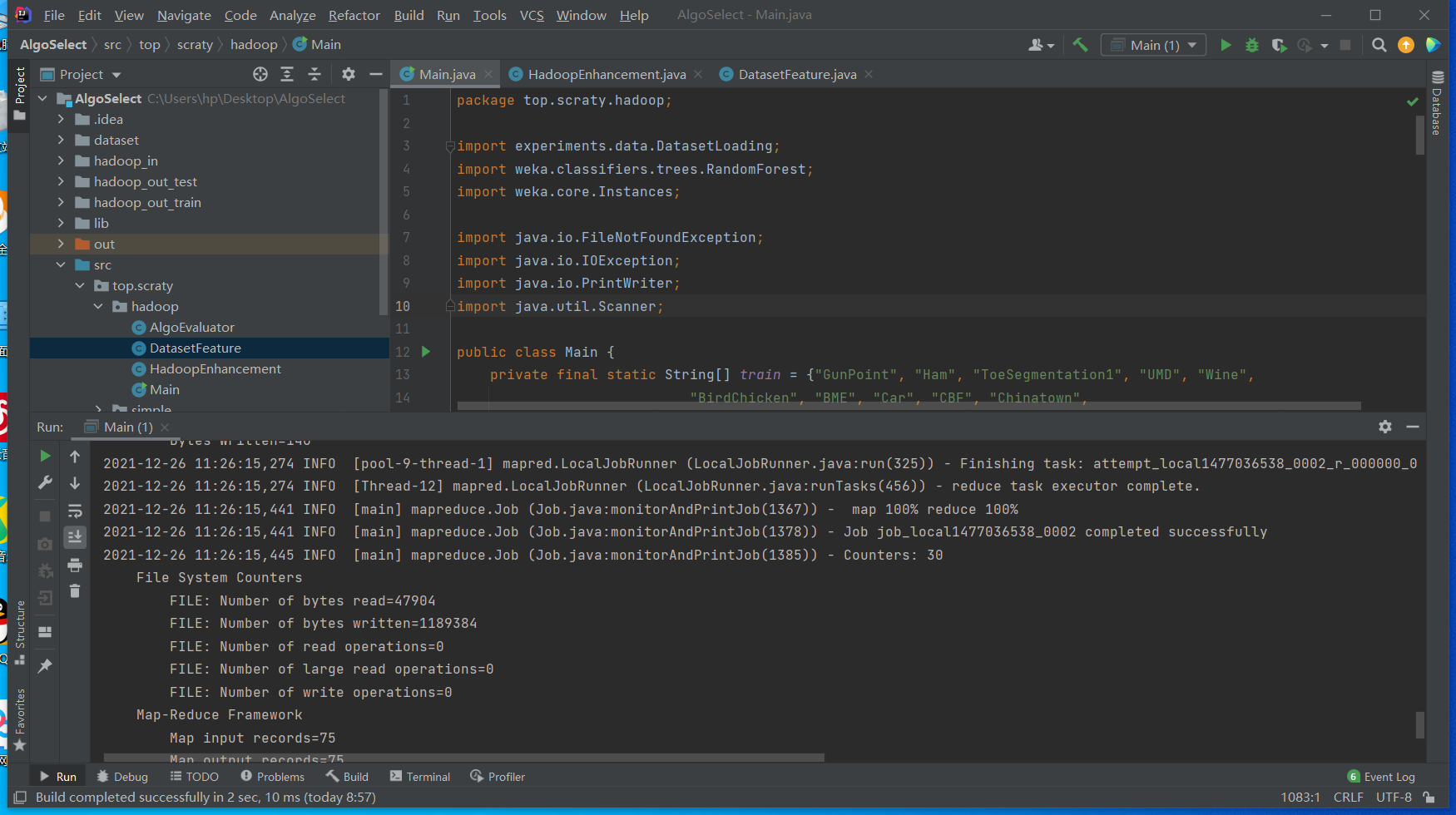
在本实验中Hadoop以本地模式运行，相信如果以伪分布式模式甚至真分布式模式运行的话，模型会跑得更快。

[实验结果]

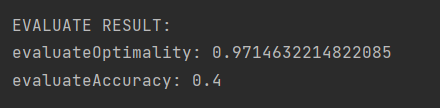
使用Hadoop加速条件下，运行时间约为两个半小时，运行截图如下：

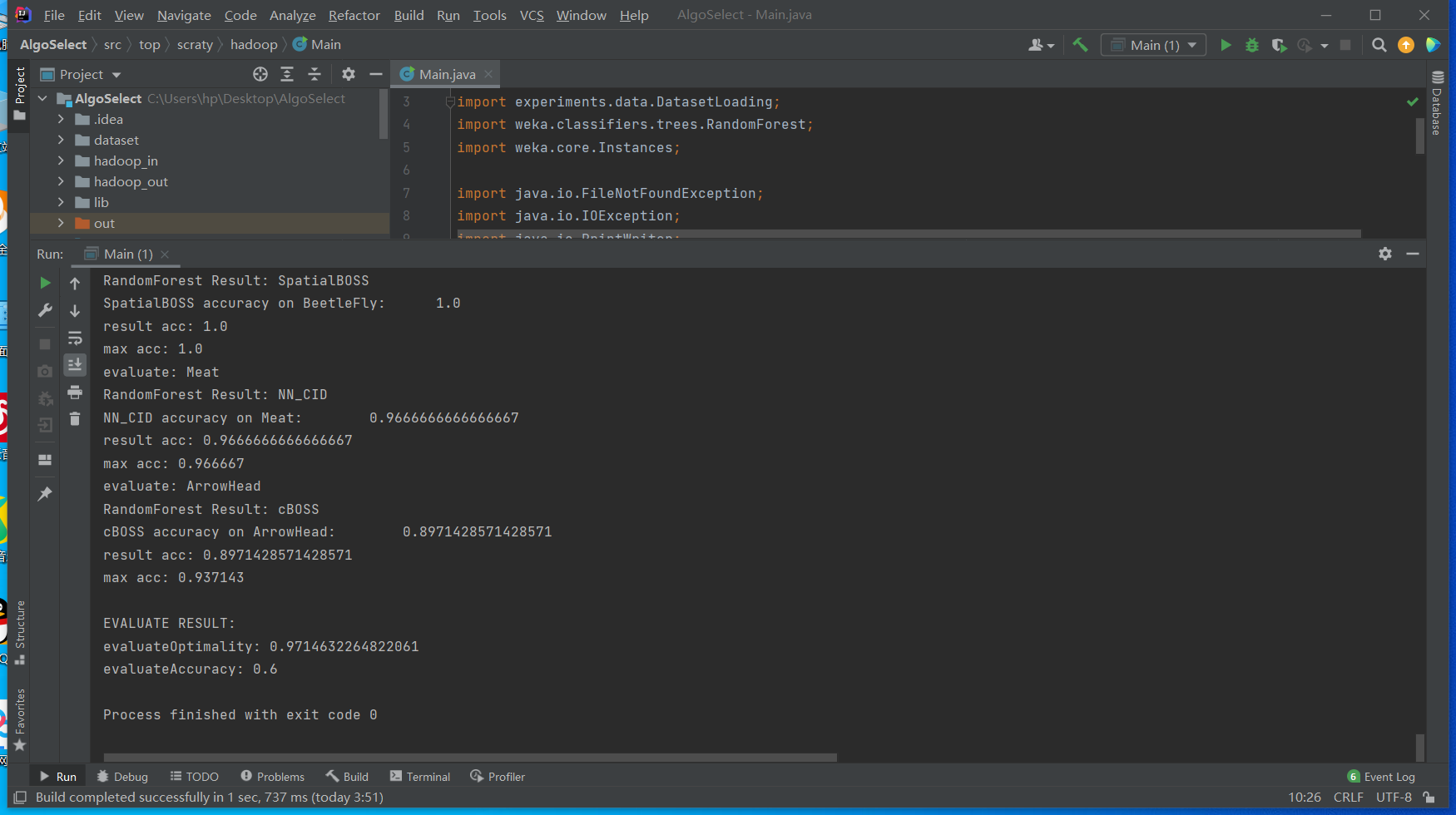






在没有Hadoop加速的条件下，运行时间约为四个小时，运行截图如下：





可以看到使用Hadoop使得运行时间得到了一定的提升。两者的运行结果基本相同，最优性均为0.97左右，准确性均为0.4。

## 第七章 相关工作

<https://arxiv.org/pdf/2007.03254.pdf> 为本文的一个相关工作的链接。在这篇论文里，作者们使用Deep Q-Network进行特征提取，并跟本文一样使用随机森林进行训练。

## 第八章 结论

本文使用两次随机投影对时间序列数据集进行特征提取，使用随机森林进行训练，并使用最优性和准确性两种指标对训练结果进行评价，初步完成了为每个数据集尽可能地选择最适合它的算法的目标。

本文还使用Hadoop加速特征提取的过程，在Map中并行地进行时间序列数据集的训练，取得了初步成效。

## 参考文献

<https://arxiv.org/pdf/2007.03254.pdf>

<https://www.timeseriesclassification.com> （本文数据集来源）

<https://github.com/uea-machine-learning/tsml> （本文调用的时间序列分类算法库）